SIMULATE PROCESS FLOW

Help Document

Sim.Pro.Flow is a decision support tool that automates the build of a discrete event simulation and allows for mapping, modelling and improving of system pathways.

This help document will describe the function of all the widgets within Sim.Pro.Flow. The document is structured where each section discusses a main tab, and the subsections discuss any subsequent tabs. This document does not provide instructions of how to use Sim.Pro.Flow but will address the considerations required by the user.

Each widget has a corresponding numbered item indicated by either a blue or purple circle, where is blue circle is standard. A purple circle indicates that a document will be produced and is described in the Output Files section.

Images outlined with a thick black dashed line indicates that this is a snippet rather than the full screen view.

The sample datasets shown were artificially generated. The dataset from Data Type 1 is used for examples shown, which consists of 16 fictional pathways that each occur once.

Note: >> indicates path, " indicates fixed name, [] indicates description of name that will change,* ***bold text* *will indicate a point in the image.**
Note: Terms defined in the glossary will be linked once on each page it occurs.

Contents

1 Usage Map **2**

2 Data Panel **3**

3 Clustering Panel **4**
3.1 Rankings and Groupings 6

4 Simulation Panel **7**
4.1 Model 7
4.1.1 Custom Inputs 8
4.2 Capacity 10
4.3 Simulation Results 11

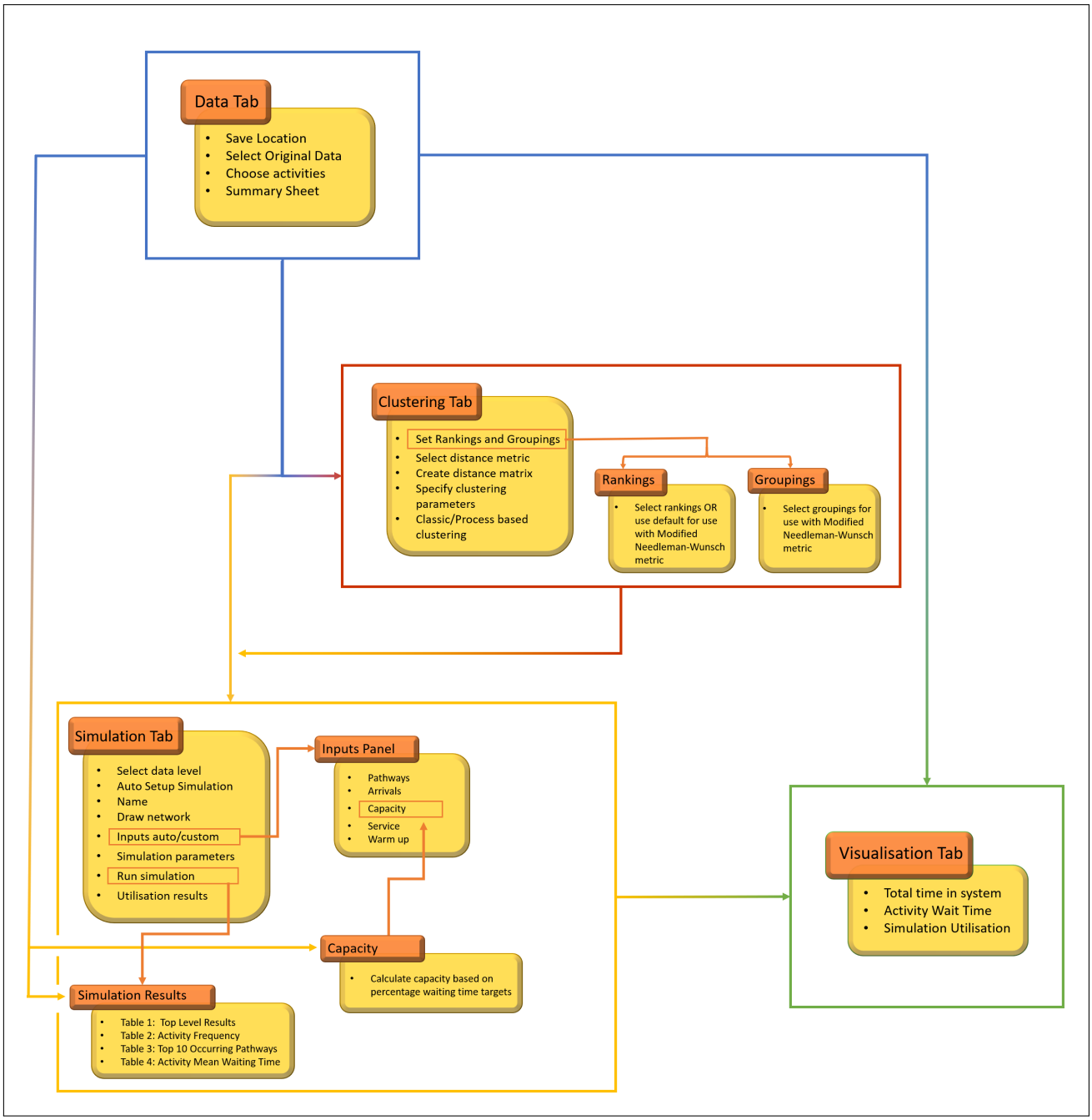
5 Visualisation Panel **12**

6 Auxiliary Information **13**
6.1 Input Data 13
6.2 Universal Widgets 15
6.3 Data Levels 15
6.4 Process Based 16
6.5 Output Files 17

7 Glossary **20**

1 Usage Map

The usage map displays the four main tabs within Sim.Pro.Flow and how information moves between them.



2 Data Panel

1) File Menu

- About - Opens the about box for Sim.Pro.Flow.
- Help - Opens this help document.
- Export - Saves the raw simulation variables.
- Save - Saves the four main results tables.

2) Select Save Location
Opens the file explorer window.
Select the folder to save the information for this session.
On selection multiple documents will be produced.
Suggested to create a new folder for each session.

3) Select Data
Opens the file explorer showing Excel files (.xlsx).
Select the dataset for this session.
Set the target days before pressing this button to use the specified target days in the Simulation Results Tab Original Results

4) Select Columns
Opens the **Select Columns pop-up box (5)** .

5) Select Columns Pop-up Box
Allows the user to select the columns required for analysis with Data Type 1 or Data Type 2.
Select all button checks all column checkboxes.
Select the **Group Codes** checkbox to indicate Data Type 2.
Only columns containing dates should be selected. Do not select blank columns.
A copy of the data will be saved in the save location, with naming convention SimProFlow_DataFileName.

6) Format
To be used with Data Type 3. Three drop down boxes will be populated with all column headers, for the user to select the corresponding column for **Unique id**, **Activity** and **Date**.
A copy of the data will be saved in the save location with naming convention 'SimProFlow_'[DataName].

7) Create Summary Sheet
Produce a word document 'Summary_Sheet.docx'.

8) Table
After the column selection has taken place (either through **(5)** or **(6)**) a non-editable table will be created displaying the activity name and the corresponding code.

3 Clustering Panel

1) Set Rankings and Groupings
Populates the Rankings and Groupings subtabs for use with the Modified Needleman-Wunsch metric.

2) Select Penalty Values
Spin control to allow user to select the penalty values for use with the Modified Needleman-Wunsch metric. Has lower and upper bounds of 0 and 100 respectively.

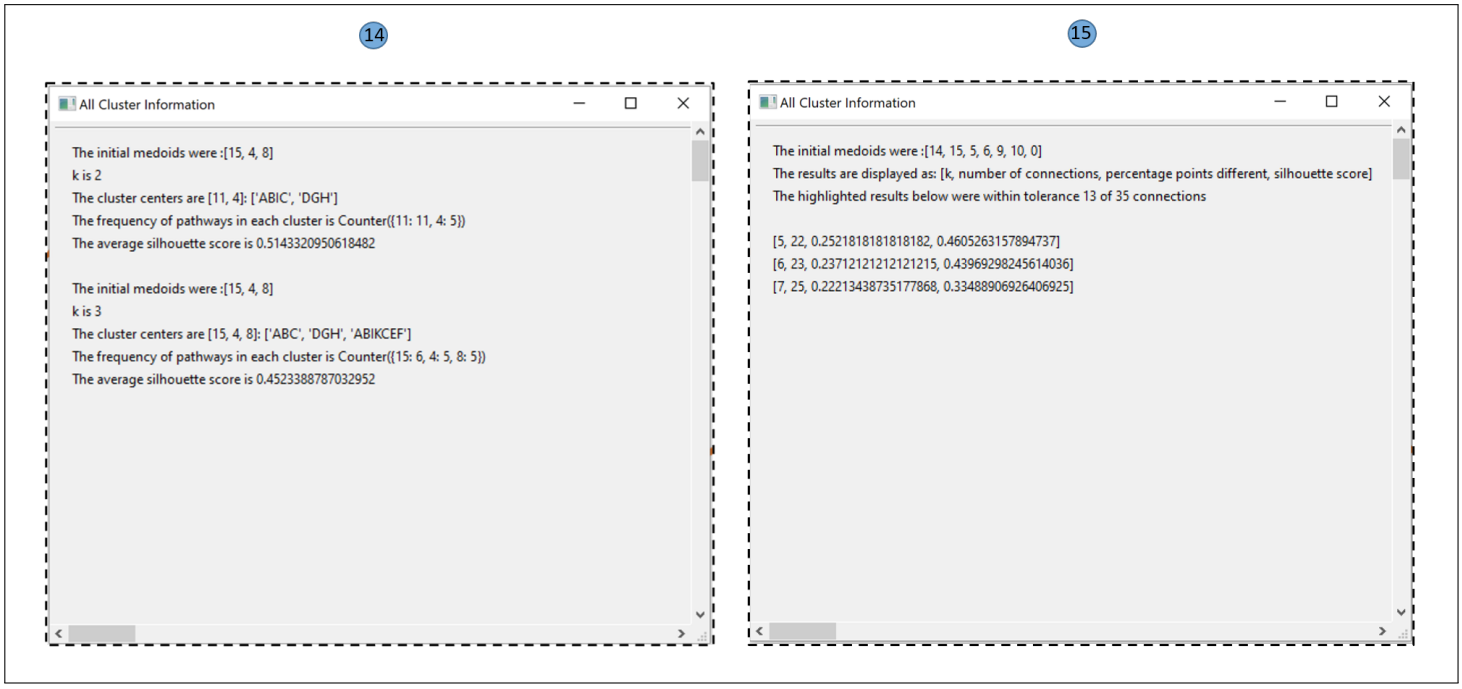
3) Select Distance Metric
Drop down box containing the eight supported distance metrics, namely: Levenshtein, Damerau-Levenshtein, Jaro, Jaro-Winkler, Needleman-Wunsch, Jaccard, Cosine, Longest Common Subsequence, and Modified Needleman-Wunsch.
**If Modified Needleman-Wunsch chosen, Rankings and Groupings must have been selected*.*

4) Create Matrix
A ixi distance matrix will be created using the chosen distance metric i.e. the distance from each pathway in the dataframe to every other pathway will be calculated.

5) Cluster Centroids
There are five methods supported for choosing the k number of initial centroids, where k is as defined in **(6)**:

- Random: Random set of unique values corresponding to pathway indexes.
- Most Occurred: Index of the pathways that occurred most often. By the definition of the dataframe, this will be the first k pathways.
- Least Distance: Pathway indexes that generated the smallest value of the sum of it's row.
- Specify: Allows the user to specify pathway indexes in the 'Specify' input box. There must be k comma separated values specified within the range [0, i].
- Previous: Select a preciously used set of initial centroids from the **Choose Previous** drop down box. This will be populated after each clustering is complete, thus can not be selected before the first clustering has occurred. The selected set must contain the same number of values as the specified k.

*. **Only one method can be selected, otherwise an error box will appear. If multiple selected post error box then the topmost selected method will be used.***



6) Cluster Results
Spin control box to select max k within range [2, 100]. Select the level of highlighted results to be displayed.

- All: Display all results in the range [2, max k]. This selection will not be able to be used with subsequent simulation.
- Besk k: Display the results for the k producing the best Silhouette Score in the range [2, max k].
- Besk k: Same as previous but k = 2 can not be selected.
- k only: Display the results for the value of k specified in the max k spin control.

7) Select Process Based Percentage
For comparing clustering results transition matrix with the adjusted transition matrix.
Selected **Adjust percentage** will be used for the adjusted transition matrix. **Used in process based clustering only.**

8) Select Process Based Tolerance
Used to specify which results to display in **(15)**.

9) Save Centroids
Saves the cluster centroids to the Excel file.
Results saved will correspond to the results choice selected in **(6).**

10) Classic Cluster (14)
k-medoids clustering applied to the distance matrix using the specifications selected in points **(5) - (9)**.
The resulting k will be used as the number of transition matrices to be used with the Clustered Transitions simulation.

11) Process Based Cluster (15)
k-medoids clustering applied to the distance matrix using the specifications selected in points **(5) - (9)**.
The resulting k will be used as the number of medoids to be used with the Process Medoids simulation.

12) Canvas
Operates as universal widget.
Displays process based violin plot to aid k selection.
Naming convention:
[Metric]_[InitialCentroidSet]_[ResultsType]_[k]

13) Toolbar
Operates as universal widget.

3.1 Rankings and Groupings

1) Select Rankings
Rankings to be used with Modified Needleman-Wunsch metric.
Drop down box to select rank for activity, where options available are [0, N].
Lower value corresponds to more important.
Choose unique value for each activity.

2) Use Default
Check box to use default rankings shown.
Default rankings calculated based on frequency of activity.

3) Confirm Rankings
Done button to confirm rankings.
This must be pressed to use Modified Needleman-Wunsch metric.

4) Select Groupings
Groupings to be used with Modified Needleman-Wunsch metric.
Drop down box to select group for activity, where options available are [Group 0, Group N].
Only activities in the same group will be able to be swapped by definition of the Modified Needleman-Wunsch metric.

5) Confirm Groupings
Done button to confirm groupings.
This must be pressed to use Modified Needleman-Wunsch metric.

4 Simulation Panel

4.1 Model

1) Select Data Level
Select the data level to use for the simulation.
Drop down options: Raw Pathways, Full Transitions, Clustered Transitions and Process Medoids.

2) Auto Setup Simulation
Construct the simulation for selected data level.
Must be initially pressed to start using this tab.
Should be pressed to change data level as selected.
****For clustered transitions and process medoids will use the most recent results. Please ensure that this was not the results type* ***'All'****.****

3) Provide Simulation Name
Name for simulation. Automatically filled on **Auto Setup Simulation** based on data level with convention [Initial]_Basic, where Initial = R, F, C, P respectively.
Must be unique for each execution of simulation - prompted by error.
If trials > 1 then name appends _Trials_[No.Runs]

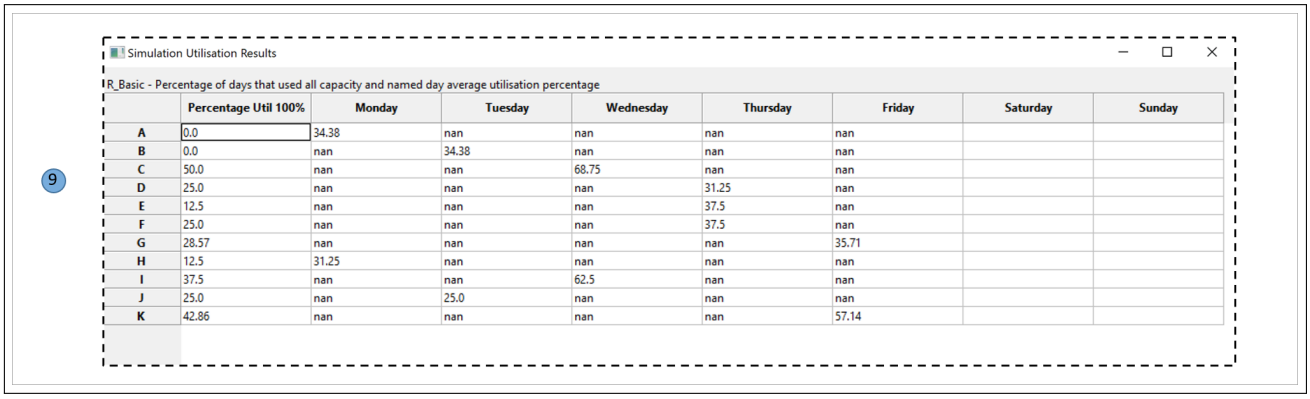
4) Network: Draw
Draw: Draw the networks for the selected data level.
Drop down box: Contains names of networks produced.
View: Open the selected network pdf in a pdf viewer.

5) Right Hand Panel
After auto setup simulation panel will populate as shown in Custom Inputs.

6) Simulation Custom Inputs
Use auto or custom values for simulation inputs.

7) Time Elements
Simulation parameters relating to time.

- Week Type: Number of working days in a week. Options 5 days, 7 days.
  Used in Simulation Custom Input - Capacity.
- Number of Individuals: Simulate until number of individuals specified have completed simulation.
  Auto - Number of individuals in data,
  Custom - defined in Arrivals.
- Target Days: Target days for time in system.
  Used in tabs Data and Simulation Results.
- Number of Trials: Number of simulation runs.
  Simulation seed set to the corresponding run number starting 0.
- Simulation Seed: Seed used for the simulation.
  **Not used if Number of Trials > 1.*

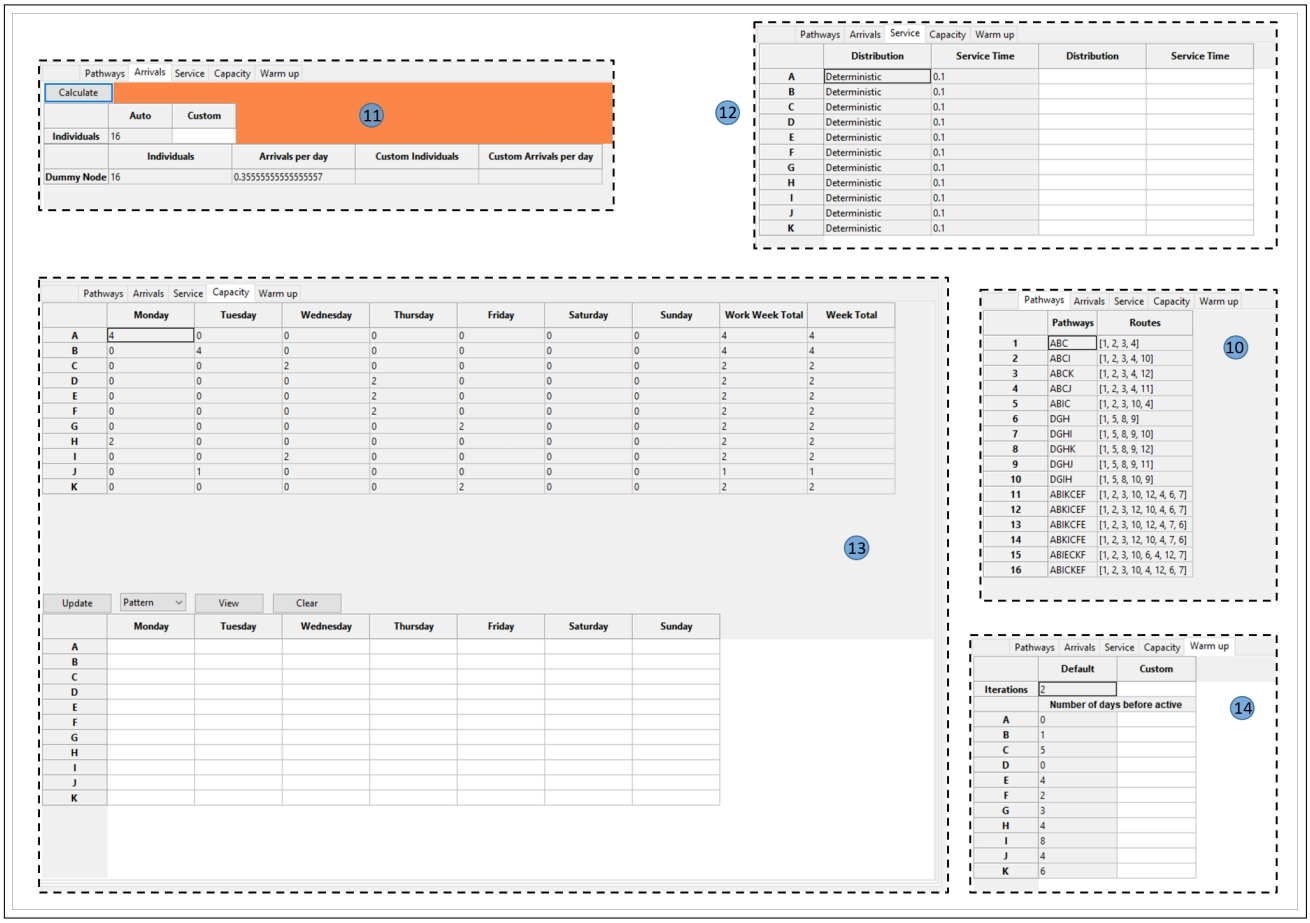


	Percentage Util 100%	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
A	0.0	34.38	nan	nan	nan	nan		
B	0.0	nan	34.38	nan	nan	nan		
C	50.0	nan	nan	68.75	nan	nan		
D	25.0	nan	nan	nan	31.25	nan		
E	12.5	nan	nan	nan	37.5	nan		
F	25.0	nan	nan	nan	37.5	nan		
G	28.57	nan	nan	nan	nan	35.71		
H	12.5	31.25	nan	nan	nan	nan		
I	37.5	nan	nan	62.5	nan	nan		
J	25.0	nan	25.0	nan	nan	nan		
K	42.86	nan	nan	nan	nan	57.14		

8) Run Simulation
Runs the simulation.
Pop up box will confirm when simulation is complete.

9) View Utilisation Results
Opens Simulation Utilisation Results pop up box.
Percentage Util 100% displays the percentage of days that used 100% of the capacity available.
Displays the mean average percentage of capacity used on each named day.
Not produced for trials.

4.1.1 Custom Inputs



	Auto	Custom
Individuals	16	

	Individuals	Arrivals per day	Custom Individuals	Custom Arrivals per day
Dummy Node	16	0.35555555555555557		

	Distribution	Service Time	Distribution	Service Time
A	Deterministic	0.1		
B	Deterministic	0.1		
C	Deterministic	0.1		
D	Deterministic	0.1		
E	Deterministic	0.1		
F	Deterministic	0.1		
G	Deterministic	0.1		
H	Deterministic	0.1		
I	Deterministic	0.1		
J	Deterministic	0.1		
K	Deterministic	0.1		

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Work Week Total	Week Total
A	4	0	0	0	0	0	0	4	4
B	0	4	0	0	0	0	0	4	4
C	0	0	2	0	0	0	0	2	2
D	0	0	0	2	0	0	0	2	2
E	0	0	0	2	0	0	0	2	2
F	0	0	0	2	0	0	0	2	2
G	0	0	0	0	2	0	0	2	2
H	2	0	0	0	0	0	0	2	2
I	0	0	2	0	0	0	0	2	2
J	0	1	0	0	0	0	0	1	1
K	0	0	0	0	2	0	0	2	2

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
A							
B							
C							
D							
E							
F							
G							
H							
I							
J							
K							

	Pathways	Routes
1	ABC	[1, 2, 3, 4]
2	ABCI	[1, 2, 3, 4, 10]
3	ABCK	[1, 2, 3, 4, 12]
4	ABCJ	[1, 2, 3, 4, 11]
5	ABIC	[1, 2, 3, 10, 4]
6	DGH	[1, 5, 8, 9]
7	DGHI	[1, 5, 8, 9, 10]
8	DGHK	[1, 5, 8, 9, 12]
9	DGHJ	[1, 5, 8, 9, 11]
10	DGIH	[1, 5, 8, 10, 9]
11	ABIKCEF	[1, 2, 3, 10, 12, 4, 6, 7]
12	ABKICEF	[1, 2, 3, 12, 10, 4, 6, 7]
13	ABIKCFE	[1, 2, 3, 10, 12, 4, 7, 6]
14	ABKICFE	[1, 2, 3, 12, 10, 4, 7, 6]
15	ABIECKF	[1, 2, 3, 10, 6, 4, 12, 7]
16	ABICKEF	[1, 2, 3, 10, 4, 12, 6, 7]

	Default	Custom
Iterations	2	
	Number of days before active	
A	0	
B	1	
C	5	
D	0	
E	4	
F	2	
G	3	
H	4	
I	8	
J	4	
K	6	

From the simulation model tab the simulation inputs **(6)** can be selected as either auto or custom. Both the auto and custom options are hosted within this right hand panel **(5)** and initially appear after auto setup simulation **(2)** has been performed. The pathways and arrivals are data level specific and shall be discussed for each, whereas the remaining inputs are universal across all data levels.
Note: The example shown are for raw pathways. Note: Any cell in grey is non editable.

1) Pathways
This displays the pathways that will be taken by the individuals.

- Raw Pathways: Table will contain two columns - Pathways and Routes.
  The pathways will correspond with the dataframe and the routes will reflect the route to take through the nodes of the pathways.
  The raw pathways all enter the simulation through a dummy node (node 1) and thus all pathways will start with 1, with the resulting nodes corresponding to letter code position +1.
- Full Transitions: This will displays a N xN transition probability matrix.
- Clustered Transitions: Takes the results from the classic clustering for k clusters and produces a NxN transition probability matrix for each.
  Each cluster will be represented by a class .
- Process Medoids: Takes the results from the process based clustering for k medoids and displays the table as is for the Raw Pathways, where the pathways correspond to the k medoids.

2) Arrivals and Number of Individuals
The arrivals correspond to the arrival rate for the simulation. The value/s for the arrivals per day will be used as the λ for an Exponential distribution. The general arrival rate is defined in Equation 1. This is only editable through the number of individuals, where the user can enter the number of individuals in the 'Custom' cell and on execution of the 'Calculate' button the arrival rate will be re-calculated. The user can then choose to either simulate until the auto or custom individuals and use the auto or custom arrivals independently (**Custom individuals not available for Raw Pathways.**).

- Raw Pathways: Arrivals will be received at a dummy node (node 1) where the individual will be assigned their pathway.
  Arrival equation applied where $A_n = \text{N}$.
  The individuals pathway number will correspond to the individuals id number.
- Full Transitions: Individuals will arrive into respective arrival activities as extracted from the transition probability matrix.
  Table will have N rows, one for each activity, with arrival equation calculated for each.
- Clustered Transitions: Same as the Full transitions with the addition that the columns will be repeated k times i.e. arrival matrix for each class.
  Arrival equation values for A_n are obtained from the propagated values.
- Process Medoids: Table will have k rows where the arrival activity will correspond to the first activity of the medoid for each class.
  Arrival equation values for A_n are obtained from the propagated values.

3) Service
Service rates are auto set to deterministic at 0.1.
Custom distributions can be deterministic or exponential. Service time can take a float value and will be formatted as necessary for the chosen distribution.

4) Capacity
Top table displays the auto calculated capacity for each activity - the average number of individuals seen on each named day .
Bottom table is used for capacity where users can manually input custom values for capacity.
Various custom options are available to auto-fill the custom table using the following widgets:

- Update: Update options in the drop down box.
- View: Populate bottom table with chosen option.
- Clear: Clear the table.

Three custom option types are available which will consider the week type (5 days or 7 days) chosen:

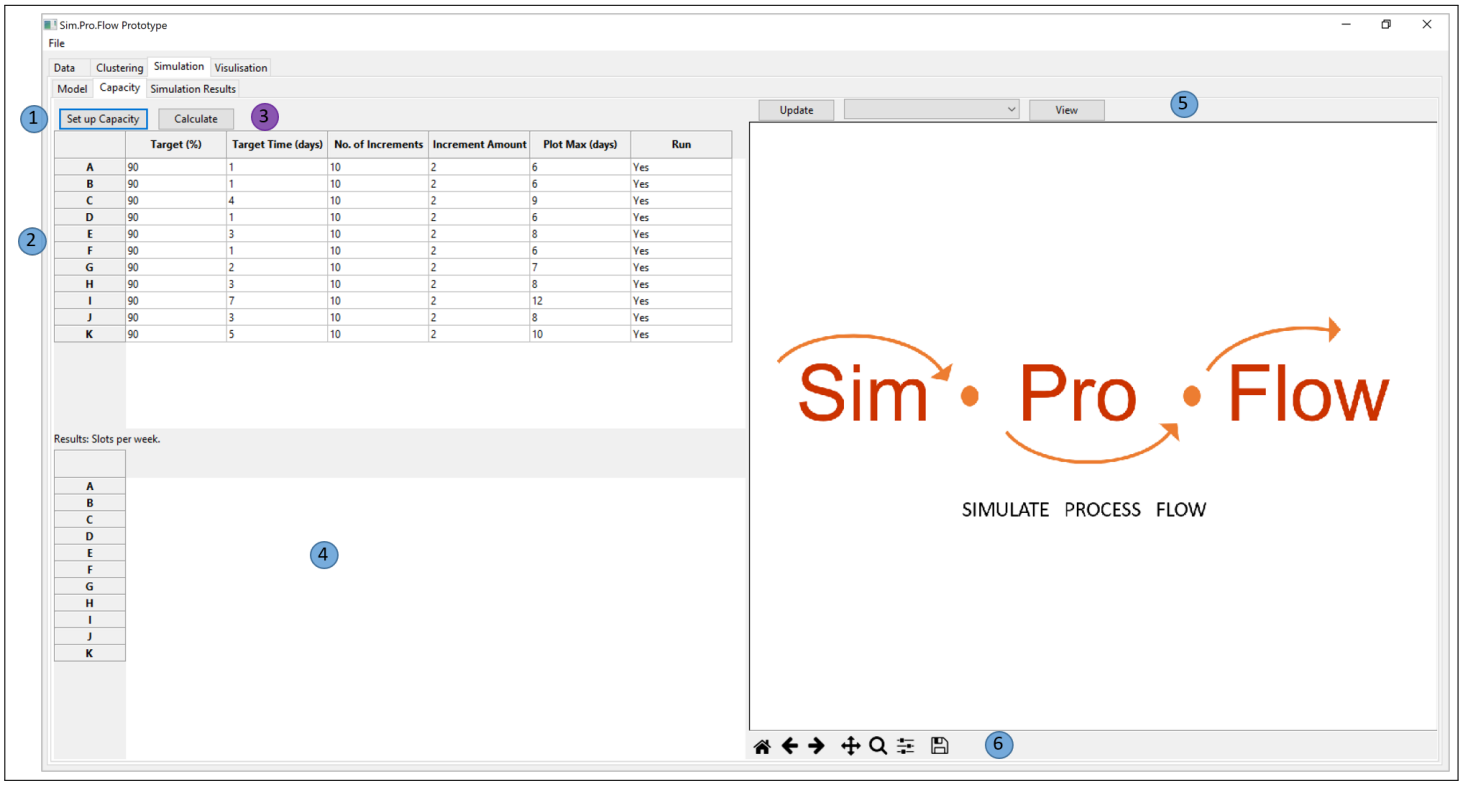
- Pattern - Use the auto pattern from the top table.
- Smoothed - Spread the total week capacity (considering chosen week type) evenly across the working days.
- Naming Convention: Cal_Cap_[Number] - Takes the results calculated in the Capacity tab and uses the smoothed technique.

Chosen capacity will create a Server Schedule

5) Warm-up
Three warm up types available:

- None: No warm up used. This is the default.
- Iterative: Collect results once w multiples of I have completed the simulation. Auto set to 2.
- Warm Start: Block service at activity for an initial number of days. Auto set to original mean activity waiting time (Table 4).

4.2 Capacity



This calculation, developed by Arruda et al [1], will suggest the amount of weekly capacity for each activity required to allow a percentage of individuals to wait no more than a target number of days e.g. 90% of individuals to wait no more than 10 days. Various values of capacity will be tested, based on the inputs provided, to suggest the number of slots per week required. The base value is auto calculated from the activity demand.

1) Set up Capacity
Create and populate capacity **Inputs Table (2)**.

2) Inputs Table
Table to control inputs for capacity calculation.

- Target (%): Percentage to wait no longer than **Target Time**.
- Target Time (days): The target number of waiting days.
  Set up Capacity will auto populate this with the floor of the average waiting time for each activity in the original data i.e. Table 4.
  The minimum value allowed is 1.
- No. of Increments: This will set the number of values to explore.
- Increment Amount: This will set the amount to increment the base value by.
- Plot Max (days): This is the maximum number of days to view on the plot.
  Auto populated to Target Time + 5.
- Run: Choose if Yes/No to run the calculation for this activity.

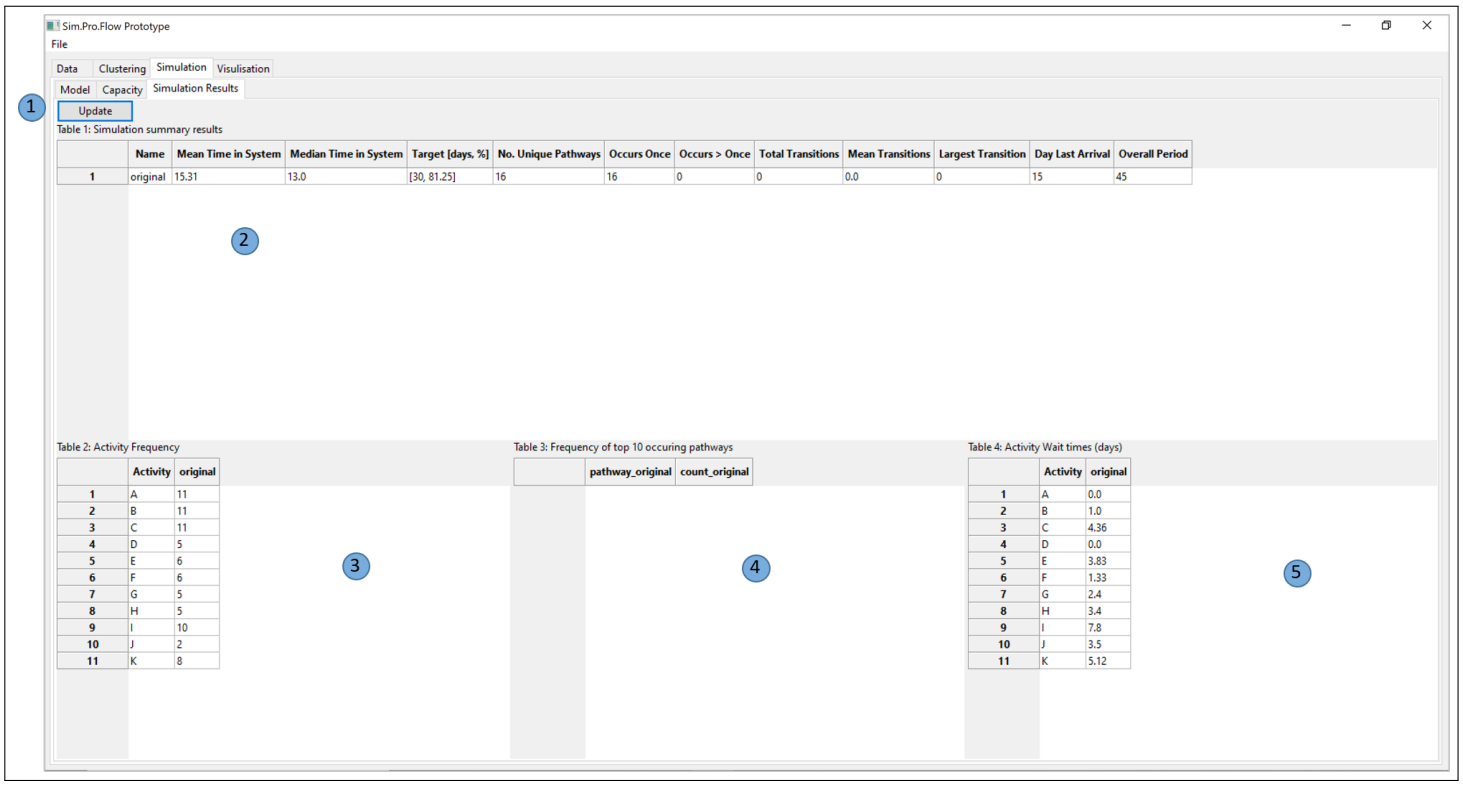
3) Calculate
This performs the calculation described in [1].
Considers Week Type selection

4) Results
A value indicating the number of capacity slots required each week for each activity will be produced.
If No selected for an activity, the previous value will be shown in the **results table (4) in red font.**
If No selected for first run then red 0 will appear in results table.

5) Canvas
Operates as universal widget.
Displays line chart showing the percentage waiting no more than x (axis value) days.
Naming convention: Cal_Cap_NumberOfCalculations

6) Toolbar
Operates as universal widget.

4.3 Simulation Results



1) Update Tables
Update button will update all tables with any new results that have been produced since the last update.
On update Table 1 will append rows, all other tables will append columns.
The first entry for all the tables will correspond to the original input data and is named original

2) Table 1
Summary Results:

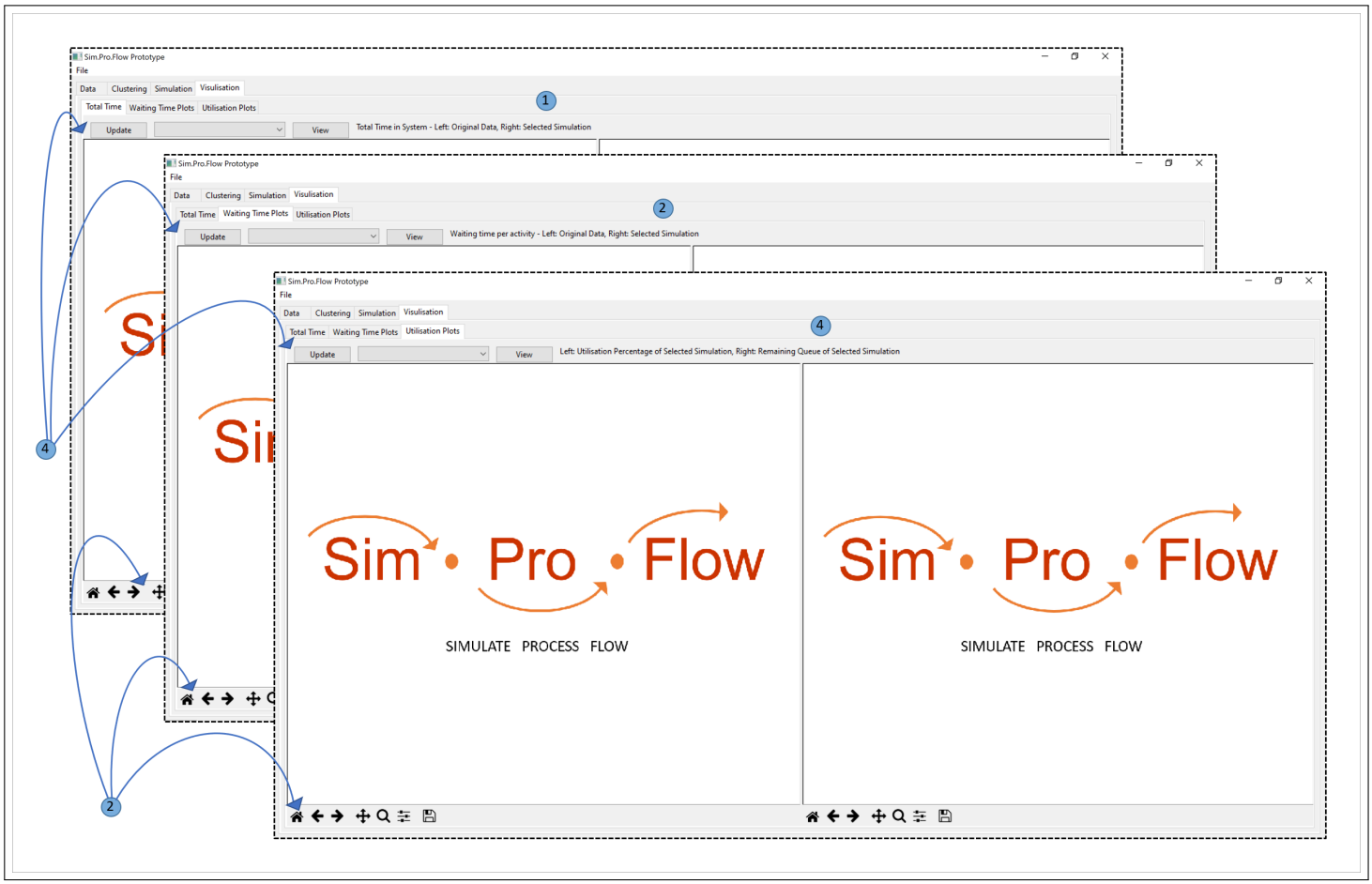
- Mean Time in System: Mean average time from arrival to exit in days.
- Median Time in System: Median average time from arrival to exit in days.
- Target [days, %]: Percentage total time in system less than the specified target days.
- No. Unique Pathways: Number of unique pathways performed.
- Occurs Once: Number of pathways that occur only once.
- Occurs > Once: Number of pathways that occur more than once.
- Total Transitions: The total sum of the difference matrix.
- Mean Transitions: The mean average of the difference matrix.
- Largest Transitions: The largest entry in the difference matrix.
- Day Last Arrival: The day that the last arrival occurred.
- Overall Period: The total number of days that occurred.

3) Table 2
Activity Frequency: The number of times each activity was performed.

4) Table 3
The top 10 occurring pathways and their corresponding frequency count.
Only pathways occurring more than once will appear.

5) Table 4
Activity Wait times (days): Mean average wait time for each activity in days.

5 Visualisation Panel



1) Total Time Tab
Displays a of histogram for total time in system, where the left canvas is for the original data and the right canvas is for the selected simulation.
The original data canvas is possible to view before any simulation has been run

2) Waiting Time Tab
Displays a plot containing multiple histograms displaying the waiting time for each activity, where the left canvas is for the original data and the right canvas is for the selected simulation.
The original data canvas is possible to view before any simulation has been run

3) Utilisation Tab
Displays two line charts of the capacity utilisation, where the left canvas plots the percentage of the capacity utilised each day and the right canvas plots the number of individuals remaining in the queue at the end of each day.

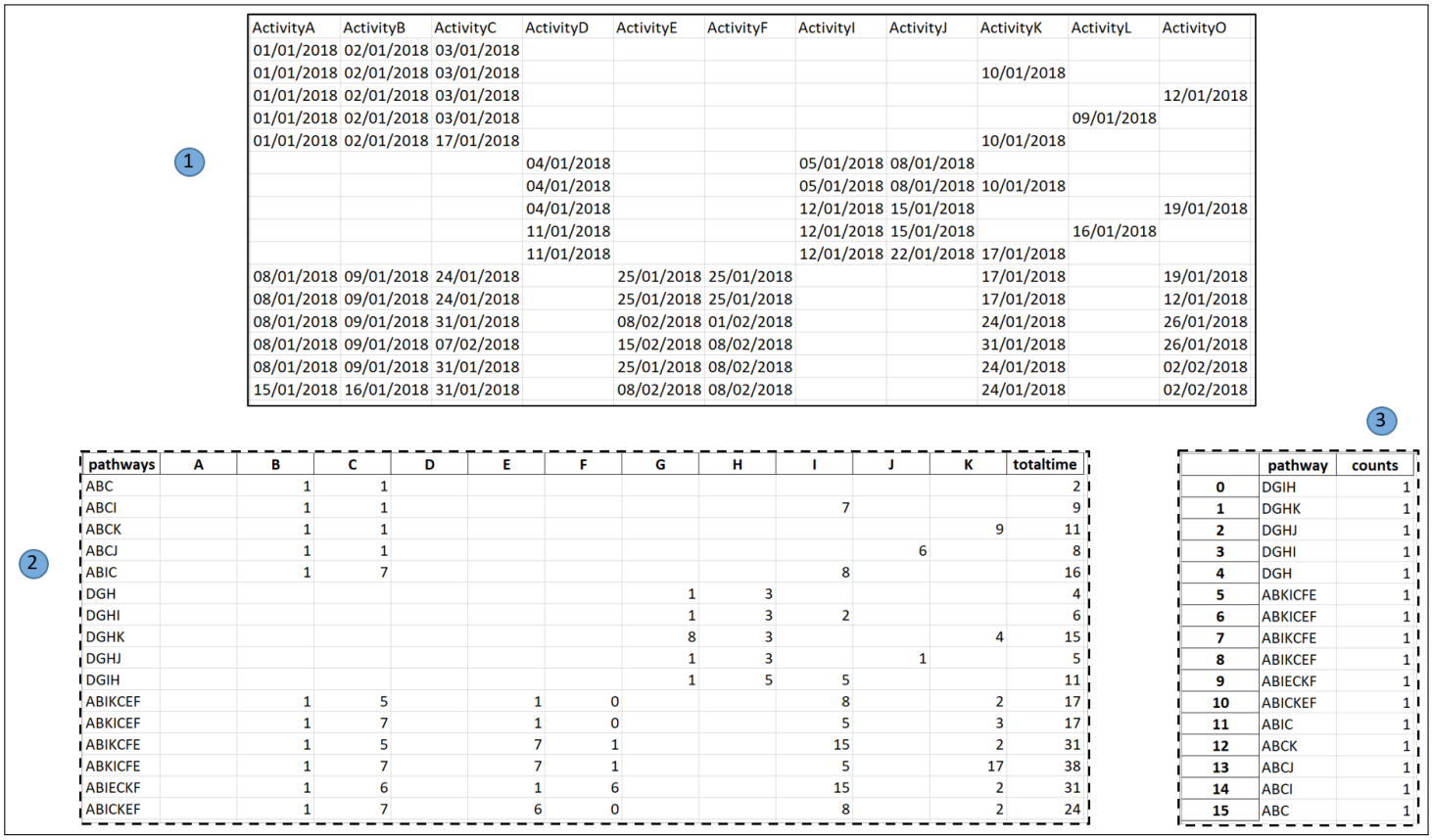
12) Canvas
Operates as universal widget.
Drop down options will be populated with simulation names.

13) Toolbar
Operates as universal widget.

*Note: *After view pressed, each canvas will required manual updating by selecting the move tool and clicking in the canvas area**
*Note: **The plot axis scale and subplot configuration is automated to best suit each plot.***
*Note: ***These plots are not available for trial simulations.****

6 Auxiliary Information

6.1 Input Data



1

ActivityA	ActivityB	ActivityC	ActivityD	ActivityE	ActivityF	ActivityI	ActivityJ	ActivityK	ActivityL	ActivityO
01/01/2018	02/01/2018	03/01/2018								
01/01/2018	02/01/2018	03/01/2018						10/01/2018		
01/01/2018	02/01/2018	03/01/2018								12/01/2018
01/01/2018	02/01/2018	03/01/2018							09/01/2018	
01/01/2018	02/01/2018	17/01/2018						10/01/2018		
			04/01/2018			05/01/2018	08/01/2018			
			04/01/2018			05/01/2018	08/01/2018	10/01/2018		
			04/01/2018			12/01/2018	15/01/2018			19/01/2018
			11/01/2018			12/01/2018	15/01/2018		16/01/2018	
			11/01/2018			12/01/2018	22/01/2018	17/01/2018		
08/01/2018	09/01/2018	24/01/2018		25/01/2018	25/01/2018			17/01/2018		19/01/2018
08/01/2018	09/01/2018	24/01/2018		25/01/2018	25/01/2018			17/01/2018		12/01/2018
08/01/2018	09/01/2018	31/01/2018		08/02/2018	01/02/2018			24/01/2018		26/01/2018
08/01/2018	09/01/2018	07/02/2018		15/02/2018	08/02/2018			31/01/2018		26/01/2018
08/01/2018	09/01/2018	31/01/2018		25/01/2018	08/02/2018			24/01/2018		02/02/2018
15/01/2018	16/01/2018	31/01/2018		08/02/2018	08/02/2018			24/01/2018		02/02/2018

2

pathways	A	B	C	D	E	F	G	H	I	J	K	totaltime
ABC		1	1									2
ABCI		1	1						7			9
ABCK		1	1								9	11
ABCJ		1	1							6		8
ABIC		1	7						8			16
DGH							1	3				4
DGHI							1	3	2			6
DGHK							8	3			4	15
DGHJ							1	3		1		5
DGIH							1	5	5			11
ABIKCEF		1	5		1	0			8		2	17
ABKICEF		1	7		1	0			5		3	17
ABIKCFE		1	5		7	1			15		2	31
ABKICFE		1	7		7	1			5		17	38
ABIECKF		1	6		1	6			15		2	31
ABICKEF		1	7		6	0			8		2	24

3

	pathway	counts
0	DGIH	1
1	DGHK	1
2	DGHJ	1
3	DGHI	1
4	DGH	1
5	ABKICFE	1
6	ABKICEF	1
7	ABIKCFE	1
8	ABIKCEF	1
9	ABIECKF	1
10	ABICKEF	1
11	ABIC	1
12	ABCK	1
13	ABCJ	1
14	ABCI	1
15	ABC	1

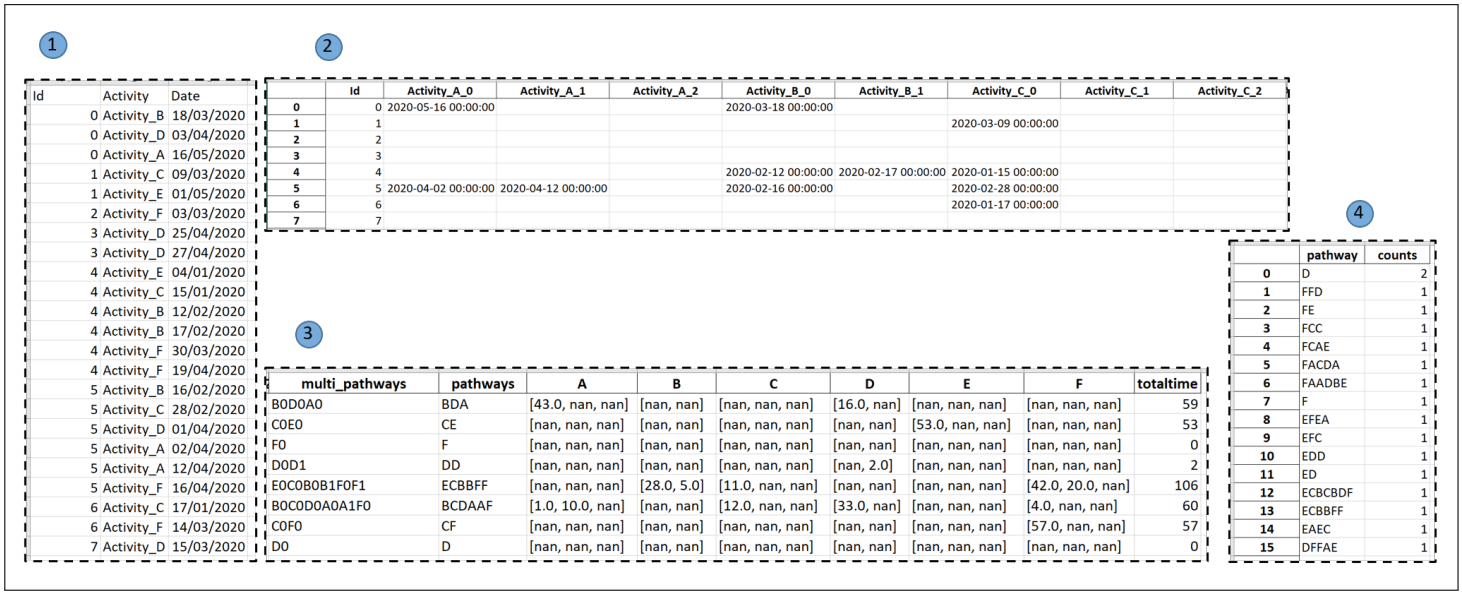
Type 1

1) Original Input Data
Each row corresponds to an individual.
Each activity has it's own column.
Date stamps to record activity and left blank otherwise.
Do not select any blank columns.

2) SimProFlow_[DataName] - Data Tab
Additional information appended to **Original Input Data** and saved in the save location as an Excel file with naming convention SimProFlow_[DataName].
Description of the columns:

- pathways: Displays the string representing the pathway taken by the individual.
- [Letter]: Waiting time for performed activity, in days.
- totaltime: This is the total waiting time in the system, in days.

3) SimProFlow_[DataName] -Dataframe Tab
This contains the unique pathways and the number of times they were performed (counts).
This will be sorted by decreasing count then alphabetically decreasing.



1

Id	Activity	Date
0	Activity_B	18/03/2020
0	Activity_D	03/04/2020
0	Activity_A	16/05/2020
1	Activity_C	09/03/2020
1	Activity_E	01/05/2020
2	Activity_F	03/03/2020
3	Activity_D	25/04/2020
3	Activity_D	27/04/2020
4	Activity_E	04/01/2020
4	Activity_C	15/01/2020
4	Activity_B	12/02/2020
4	Activity_B	17/02/2020
4	Activity_F	30/03/2020
4	Activity_F	19/04/2020
5	Activity_B	16/02/2020
5	Activity_C	28/02/2020
5	Activity_D	01/04/2020
5	Activity_A	02/04/2020
5	Activity_A	12/04/2020
5	Activity_F	16/04/2020
6	Activity_C	17/01/2020
6	Activity_F	14/03/2020
7	Activity_D	15/03/2020

2

	Id	Activity_A_0	Activity_A_1	Activity_A_2	Activity_B_0	Activity_B_1	Activity_C_0	Activity_C_1	Activity_C_2
0	0	2020-05-16 00:00:00			2020-03-18 00:00:00				
1	1						2020-03-09 00:00:00		
2	2								
3	3								
4	4				2020-02-12 00:00:00	2020-02-17 00:00:00	2020-01-15 00:00:00		
5	5	2020-04-02 00:00:00	2020-04-12 00:00:00		2020-02-16 00:00:00		2020-02-28 00:00:00		
6	6						2020-01-17 00:00:00		
7	7								

3

multi_pathways	pathways	A	B	C	D	E	F	totaltime
B0D0A0	BDA	[43.0, nan, nan]	[nan, nan]	[nan, nan, nan]	[16.0, nan]	[nan, nan, nan]	[nan, nan, nan]	59
C0E0	CE	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[nan, nan]	[53.0, nan, nan]	[nan, nan, nan]	53
F0	F	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[nan, nan, nan]	0
D0D1	DD	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[nan, 2.0]	[nan, nan, nan]	[nan, nan, nan]	2
E0C0B0B1F0F1	ECBBFF	[nan, nan, nan]	[28.0, 5.0]	[11.0, nan, nan]	[nan, nan]	[nan, nan, nan]	[42.0, 20.0, nan]	106
B0C0D0A0A1F0	BCDAAF	[1.0, 10.0, nan]	[nan, nan]	[12.0, nan, nan]	[33.0, nan]	[nan, nan, nan]	[4.0, nan, nan]	60
C0F0	CF	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[57.0, nan, nan]	57
D0	D	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[nan, nan]	[nan, nan, nan]	[nan, nan, nan]	0

4

	pathway	counts
0	D	2
1	FFD	1
2	FE	1
3	FCC	1
4	FCAE	1
5	FACDA	1
6	FAADBE	1
7	F	1
8	EFEA	1
9	EFC	1
10	EDD	1
11	ED	1
12	ECBCBDF	1
13	ECBBFF	1
14	EAEC	1
15	DFFAE	1

Type 2

1) Original Input Data
Three columns:

- Id: Containing unique identifier for an individual. **This can be alphanumeric.**
- Activity: Containing activity names.
- Date: Containing performance date.

Columns do not have to have these specific titles.* **Dataset can contain additional (redundant) columns.*

2) SimProFlow_[DataName] - Data Tab
Converted **Original Input Data** into the format where each row corresponds to an individual and each column corresponds to a performance of an activity. [Activity name] followed by [_Number] corresponds to the performance occurrence number, starting at 0 e.g. _0 is the first time, _1 is the second time.
This should correctly correlate to lower [_Number] and earlier date.

3) SimProFlow_[DataName] - Data Tab Cont.
Additional information appended to **Original Input Data** and saved in the save location as an Excel file with naming convention SimProFlow_[DataName].
Description of the columns:

- multi_pathways: This will contain the pathway string consisting of the letter code followed by the [_Number].
- pathways: Displays the string representing the pathway taken by the individual.
- [Letter]: Waiting time for performed activity, in days.
  This will be a list of values the length of the maximum number of times one individual performed that activity e.g. Column D contains lists of length 2 as pathway 3 (D0D1) contains two performances of activity D.
  nan corresponds to either the activity wasn't performed or it was the first activity and doesn't have a waiting time e.g. pathway 1 (C0E0) has list [nan, nan, nan] for C even though C was performed, as this was the first activity.
- totaltime: This is the total waiting time in the system, in days.

4) SimProFlow_[DataName] - Dataframe Tab
This contains the unique pathways and the number of times they were performed (counts).
This will be sorted by decreasing count then alphabetically decreasing.

Type 3

This will take in data of the form **(2)** and group together activities with the same name but differing [_Number]. For example Activity_A_0, Activity_A_1 and Activity_A_2 will be considered the same activity.

6.2 Universal Widgets

Two widgets are repeatedly used throughout the interface.

Canvas
The panel is made up of:

- Update button: Update the options available in the drop down box.
- Drop Down box: Show the options available to be viewed.
- View: Load the selected option into the canvas. **The option will not appear on the canvas until it has been activated via a* ***Toolbar*** *item.**

All options available to view will be saved in a corresponding location within the save location and called upon to view.
The Sim.Pro.Flow logo is used as a placeholder.

Toolbar
This is matplotlib's interactive navigation toolbar [13]. The two main tools allow the user to move around the canvas and zoom into a selected area.
To move: Select the cross arrow to click and drag around the image.
This item can be used to activate the canvas.
To zoom: Select the magnifying glass to click and drag to form a box which will enlarge to fill the canvas.
The home tool will reset the image and the back and forward arrows will move between the previous and next actions performed respectively.
It is not necessary to use the save tool as all images will have been saved in the save location.

6.3 Data Levels

There are four possible data levels to be used within Sim.Pro.Flow. Each data level explores a deeper granular level for mapping the pathways. Each data level is described in turn, including the advantages and disadvantages, to help the user decide the most appropriate data level to use.

- **Raw Pathways:** These are the pathways exactly as they appear in the data. The purpose of using the raw pathways is to allow validation of the simulation parameters, by ensuring that the key performance indicators (KPI's) are reflective of those seen in the data. This is a major advantage and eases the user in to a system exploring pathways that are familiar and have actually previously happened in reality. The main disadvantage of this level is that it is very specific to the input data and doesn't allow for any variation.

- **Full Transitions:** The transition matrix can be extracted from the raw pathways. This allows for some variation to be introduced into system, which can be more reflective of future events. The disadvantages are that this can lead to unrealistic and impossible pathways. As the movements through the system will be based off probabilities extracted from the data, it could lead to individuals bouncing around within the system and performing really long pathways. Furthermore, restricting an activity to be performed at most once (referred to as at most once restriction) can not be enforced here.

- **Cluster Transitions:** From the results of the clustering the transition matrix for each cluster can be extracted. The aim for this level is to have more precise transition matrices, as the pathways have been clustered with those that they are most similar to, and thus resulting pathways should display less variation than the full transitions. This method has all the same disadvantages as the Full Transitions.

- **Process Medoids:** The second piece of usable information that is produced from the clustering are the pathways that are selected as the medoids. These exact pathways can be performed, and as such solves the disadvantages of using the transition matrices, as now only previous pathways can be performed and if the input data has the at most once restriction, so will the process medoids. Similar to the Raw Pathways, this does pose the lack of variation disadvantage but now with even more restriction.

6.4 Process Based

The process based clustering will perform the same process as the class clustering (using k-medoids), however the results of the clustering will be reported and used differently. The process based will use only the pathways selected as the medoids for each cluster. Therefore extra information is provided to support the user to select the number of k clusters to use.
This will report values to represent how closely the medoids reflect the original pathways. Putting this in terms of the network, this would translate to minimise the number of edges whilst also minimising the differences in the edge values in comparison to the full network.

Selection
The calculation process is as follows:

1. In the original data matrix: Remove transition occurring less than a specified number of times,
2. Calculate this adjusted transition matrix,
3. From the medoids, construct the transition matrix, accounting for connections representing the number of pathways in the cluster,
4. Calculate the absolute difference matrix,
5. Count the number of non-zero connections,
6. Calculate the average percentage points different,
7. Plot all non-zero differences.

6.5 Output Files

Note: >> indicates path, " indicates fixed name, [] indicates description of name that will change.
Note: Main location is the selected save location folder.

Action Button	Name	Description
	Data Tab	
Select Save Location	Folder 'Network_diagrams'	Empty folder for Draw to output.
	Folder 'Plots	General folder to contain place specific folders for plots.
	'Plots' >> 'Capacity' Folder	Empty folder to contain plots from the Capacity tab.
	'Plots' >> 'Process_Violin_Plots' Folder	Empty folder to contain the process violin plots from the clustering tab.
	'Plots' >> 'Simulation' Folder	Empty folder to contain all plots from the simulation tab.
	'Plots' >> 'Trials' Folder	Empty folder to contain plots from running trials in the simulation tab.
	'Plots' >> 'Summary' Folder	Empty folder to contain the plots for the summary sheet.
	'Clustering_Transition_Matrix.xlsx'	Empty excel file to add sheets for the clustering tansition matrix from the clustering tab.
	'Process_Centroids.xlsx'	Empty excel file to add sheets for the process based clustering centroids from the clustering tab.
	'Raw_Sim_Results.xlsx'	Empty excel file to add sheets for the raw simulation results from the simulation tab.
	'Simulation_Difference_Matrix.xlsx'	Empty excel file to add sheets for the simulation difference matrix from the simulation tab.
	'Cluster_Centroids.xlsx'	Empty excel file to add sheets for the classic clustering centroids from the clustering tab.
Select Columns/ Format	'SimProFlow_[DataName].xlsx'	Additional information added to a copy of the original data file, as shown in data types.
	'Plots' >> 'Simulation' >> 'Activty_Waits_original.png'	Plot of subplots containing histogram of the waiting time for each activity in the oringal data selection.
	'Plots' >> 'Simulation' >> 'TotalTime_original.png'	Histogram of the total time in system for the original data.
Create Summary Sheet	'Summary_Sheet.docx'	Word document produced containing summary information about the original data. Some of the information included is the number of individuals, number of pathways, mean, median and quartile time in system, as well as the following four plots.
	'Plots' >> 'Summary' >> 'Activity_Frequency.png'	Horizontal bar chart displaying the number of times (frequency) that each activity was performed.
	'Plots' >> 'Summary' >> 'Boxplot_Activity_Wait_Time.png'	Boxplot of the activity waiting times.
	'Plots' >> 'Summary' >> 'Heatmap_All_Pathways.png'	Heatmap displaying a representation of all pathways from the original data.
	'Plots'>> 'Summary' >> 'Histogram_Total_Time_in_System.png'	Histogram of the total time in system. **This may be a different scale to 'TotalTime_original'.**

Action Button	Name	Description
	Clustering Tab	
Create Matrix	Update 'Clustering_Transition_Matrix.xlsx'	Sheet added with distance metric code as sheet name. If Modified Needleman-Wunsch selected sheet name is MNW_[mgsns] where m, g, s and ns are the penalty values selected. Contains nxn distance matrix for pathways in same order as dataframe.
Save Centroids Checked	Classic Cluster Used: Update 'Cluster_Centroids.xlsx'	On first save create sheet Set_Medoids recording the set number and the initial centroids. This will subsequently be updated. Sheet added called [Metric]_Set_[SetNumber]_df containing all the pathways in the dataframe and the index of the corresponding medoids of the cluster it belongs to.
	Process Cluster Used: Update 'Process_Centroids.xlsx'	On first save create sheet Set_Medoids recording the set number and the initial centroids. This will subsequently be updated. Sheet added called [Metric]_Set_[SetNumber] will contain the medoids for k (based on results type) and the number of pathways assigned to each medoids. Sheet added called [Metric]_Set_[SetNumber]_df containing all the pathways in the dataframe and the index of the corresponding medoids of the cluster it belongs to.
Process Cluster	'Plots' >> 'Process_Violin_Plots' >> [Metric_Set_SetNumber_Type_k].png	Creates the violin plot for all values in [2, max k] to aid in decision of k values to use.
	Simulation Tab	
Run Simulation	Update 'Raw_Sim_Results.xlsx'	Adds a sheet called [SimName] containing the raw simulation results. Each row is an individual with columns for id, waiting time at each activity, pathway, total time in system and customer class. Add a sheet called [SimName]_Util containing the Utilisation Table.
	Update 'Simulation_Difference_Matrix.xlsx'	Initially adds sheet called original. For each simulation adds a sheet called [SimName] containing the transition matrix excluding the start transitions and including the end transitions.
	'Plots' >> 'Simulation' >> 'Activity_Waits_'[SimName].png	Plot contianing subplots of histograms for the waiting time for each activity.
	'Plots' >> 'Simulation' >> 'TotalTime_'[SimName].png	Histogram of the total time in system.
	'Plots' >> 'Simulation' >> 'Utilisation_Percent_'[SimName].png	Plot containing subplots of a line chart showing the percentage of capacity used each day for each activity.
	'Plots' >> 'Simulation' >> 'Utilisation_Queue_'[SimName].png	Plot containing subplots of a line chart showing the number of individuals remaining in the queue at the end of each day.

Action Button	Name	Description
Note in all draw diagrams a light grey line with no label represents a value of 1, whether that be one connection or a probability of 1.		
Auto Setup - Full Transitions - Draw	'Network_diagrams' >> 'Network_'[SimName].png	Directed graph of the full transitions network. Each node is an activity where an edge between two nodes represents the transition probability. The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
Auto Setup - Clustered Transitions - Draw	'Network_diagrams' >> 'Network_'[SimName]'_Class'[_No.].png	Directed graph for each class for the clustered transitions. There will be a file created for each class/cluster where [_No.] is the class number. Each node is an activity where an edge between two nodes represents the transition probability. The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
Auto Setup - Process Based - Draw	'Network_diagrams' >> 'Network_'[SimName_k_].png	Directed graph for the process medoids. Each node is an activity where an edge between two nodes represents the transition probability from within the set of medoids. The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
	'Network_diagrams' >> 'Network_'[SimName_k_]'_adjust'[_Perc].png	Directed graph for the process medoids. Each node is an activity where an edge between two nodes represents the transition probability for the original transitions after adjusting for the adjust percentage ([_Perc]). The start and end nodes are included for visual aid. The start node itself is not the arrival node for the simulation.
	'Network_diagrams' >> 'Network_'[SimName_k_]'_pathways'.png	Visual representation of the raw medoids.
	'Network_diagrams' >> 'Network_'[SimName_k_]'_linked'.png	Visual representation of the raw medoids where activities are grouped by position. The boxes define the position of the activity (reading left to right). The boxes contain nodes representing all activities that occurred at that position in the set of medoids. The connections between activities in the groups represent the number of times that connection occurred in the set of medoids.

Simulation - Capacity Tab

Calculate	'Plots' >> 'Capacity' >> 'Cal_Cap_'[Number].png	Plot containing subplots for each activity displaying the percentage seen within x axis days for the various values of weekly capacity investigated.

File Menu

Export	'Raw_Variables.py'	Python file containing a dictionary for each of the arrivals, service, capacity and service options used for each simulation. The dictionary keys are the [SimName].
Save	'Results_Tables.xlsx'	The four results tables each saved as a sheet. This will be overwritten on each save.

7 Glossary

Term	Description
General	
Data	The selected input data.
I	The number of individuals in the data.
Dataframe	The dataset generated of the unique pathways within the data. This will be sorted by decreasing count then alphabetically decreasing.
i	The number of entries in the dataframe which corresponds to the number of unique pathways in the data.
Pathway	String of single character codes representing the chronological order of activities performed by an individual.
Pathways Indexes	The number corresponding to the position of the pathway in the dataframe, starting at 0.
Day	The time unit of a day.
Named Day	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday. Day 0 will be Monday in the simulation.
N	Number of activities.
A_n	Number of arrivals at node n i.e. the number of times node n was the first activity in the pathway.
P	Number of days in the period covered in the data.
Floor	Formally the largest integer no larger than x [28] i.e. the integer.
Ceiling	Formally the smallest integer no smaller than x [28] i.e. the integer + 1.
Networks	Depiction of the Transition Matrix in the form of a Directed Graph.
Time in System	Integer of the last service start date minus the integer of the arrival date.
Metrics	
Distance Metrics (code)	Methods used to calculate a numerical value to represent how different two strings are. The Python library Textdistance [22] is used to calculate all metrics except for the Modified Needleman-Wunsch [2].
Levenshtein (Lev)	[33, 17]
Damerau-Levenshtein (DLev)	[25, 10]
Jaro (Jaro)	[31, 14, 15]
Jaro-Winkler (JaroW)	[31, 16]
Needleman-Wunsch (NW)	[36, 18]
Jaccard (Jac)	[30]
Cosine (Cos)	[24, 9]
Longest Common Subsequence (LCS)	[35]
Modified Needleman-Wunsch (MNW)	[2]

Term	Description
Clustering	
Transition Matrix	Matrix where position (i, j) represents the exact number of times activity i followed by activity j was performed.
Adjusted Transition Matrix	Same as Transition Matrix but where any entries with values less than the Adjust Percentage are reduced to 0.
Adjust Percentage	Percentage applied to the number of unique pathways e.g. 5% of 200 pathways = 10.
Transition Probability Matrix	Matrix where position (i, j) represents the probability of activity i followed by activity j was performed. Calculated by each value in the Transition Matrix divided by the sum of its row.
Difference Matrix	The total, mean and largest transitions are calculated by comparing the original transition matrix with the simulation transition matrix, not including the start and end activities.
Distance Matrix	Matrix where position (i, j) represents the distance between string i and j.
Propagated Values	As the clustering is performed for the dataframe (only unique pathways), when using the clustering results for the simulation one must consider 'propagating' the results to reflect the full data. For example, if pathway 'ABC' was assigned to cluster 0 and repeated 4 times in the original data, then the list of pathways for cluster 0 would contain ['ABC', 'ABC', 'ABC', 'ABC'].
k-Medoids	Clustering method to separate the set of pathways into k clusters with one of the pathways as the centre point for each cluster. Points are assigned to the cluster with the smallest distance to the centre point (medoid). [32]
Medoids	The pathway selected as the centre point for a cluster.
k	The number of clusters.
Max k	The maximum number of clusters to calculate.
Initial Centroids	Initial starting centres for the k-medoids clustering.
Silhouette Score	A metric to represent how 'good' the clustering is - used to suggest the number of clusters (k) to select. Describes how a point is to its assigned cluster centre compared to the centres of the other clusters [37]. Values range between -1 and 1 where a larger values indicates a better clustering. The Python library scikit-learn is used to calculate the silhouette score [20].
Network Connections	Number of non-zero entries in the Transition Matrix
Tolerance	Value to define the results to highlight i.e. if the number of network connections are within the selected tolerance of the number of network connections of the original data then results will be highlighted for that value of k.

Term	Description
	Simulation
Seed	As Ciw [4] uses random number generators, a seed is set to ensure that the same set of random numbers are used. This allows the simulation to be run multiples times and not have the results change due to the random numbers generated.
Node	A node is an object where individuals are serviced. Each activity will have a corresponding node. Node count starts at 1.
Dummy Node	A dummy node is used when the service at the node is unimportant but the node is required to structure the system.
Class	A class is used to denote groups of individuals [5].
Arrival Rate	The rate at which individuals will arrive into the system through the arrival node/s. An Exponential Distribution is use where the general arrival rate used is calculated by $$\lambda = \frac{A_n}{P} \quad (1)$$
Exponential Distribution	The exponential distribution is used the sample the time between events which occur at an average rate λ [6, 27]. The exponential distribution is commonly used for the arrival rate.
Deterministic Distribution	The deterministic distribution will always sample the same value [6].
Server Schedule	Takes the chosen capacity and produces a schedule to define the amount of capacity required per day. The length of the schedule will be defined for the number of weeks in 1.5* Overall Period for the original data. If warm up Iterative selected then defined for the number of weeks in w *Overall Period for the original data. Schedule is cyclical, thus if the end is reached the schedule will loop. **Ensure for warm up Warm Start that the schedule is not required to loop**
Warm Up	A warm up time is the time where the simulation will run without collecting results. This is to allow the simulation to 'fill up' before observation of the system. [8]
w	For the iterative warm up, w is the number of times to simulate I individuals where only the final I individuals will record results e.g. with w = 3 and I = 200, the simulation will run for 600 individuals where only the final 200 (from id number 401 onwards) will be recorded.
Week Type	Number of working days. If **5 days** selected simulation capacity will be set to 0 for Saturday and Sunday for all activities (including the dummy node).
Percentage Util 100%	$$\text{Percentage Util } 100\% = \frac{\text{No. days recorded all capacity used}}{\text{No. days activity had capacity}} \quad (2)$$
Overall Period	For the original data this is calculated by latest date recorded - earliest date recorded. For the Simulation results this is calculated by latest exit date - earliest exit date recorded.

Term	Description
Plots	
Bar Chart	A bar chart is used to represent categorical data [23, 12].
Directed Graph	A graph consisting of vertices and directed edges [26]. The directed graph represents the network of the activity interactions in the pathways. A vertex represents a node/activity and a directed edge represents the precedence of the vertex. The python library Graphviz [11] was used to create these Directed Graphs.
Histogram	A plot of the approximate distribution for numerical data [29]. Represents the frequency of occurrences. [12]
Line Chart	A series of data points joined together with a line [34, 12].
Violin Plot	Displays the distribution of the data highlighting the mean. [38, 12, 21]

References

[1] E.F., Arruda, P. Harper, T. England, D. Gartner, E. Aspland, F.O, Ourique, T. Crosby. Resource optimization for cancer pathways with aggregate diagnostic demand: a perishable inventory approach *IMA Journal of Management Mathematics*, 2020.

[2] E. Aspland, P.R. Harper, D. Gartner, P. Webb and P. Barrett-Lee. Modified Needleman Wunsch Algorithm for Clincal Pathway Clustering. *Journal of Biomedical Informatics*, 2021.

[3] Andrei Novikov. PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230, April 2019.

[4] The Ciw library developers Ciw: 2.0.1, 2019 `https://github.com/CiwPython/Ciw`

[5] Ciw Class, The Ciw library developers Ciw: 2.0.1, 2019 `https://ciw.readthedocs.io/en/latest/Tutorial-II/tutorial_vii.html`

[6] Ciw Distributions, The Ciw library developers Ciw: 2.0.1, 2019 `https://ciw.readthedocs.io/en/latest/Guides/set_distributions.html`

[7] Ciw seed, The Ciw library developers Ciw: 2.0.1, 2019 `https://ciw.readthedocs.io/en/latest/Guides/seed.html`

[8] Ciw Warm up, The Ciw library developers Ciw: 2.0.1, 2019 `https://ciw.readthedocs.io/en/latest/Tutorial-I/tutorial_iv.html`

[9] P-N. Tan, M. Steinbach and V. Kumar. Introduction to Data Mining. Chapter 8, page 500, 2005.

[10] F.J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, Vol 7, Issue 3:171–176, 1964.

[11] Graphviz, Python library, `https://graphviz.readthedocs.io/en/stable/index.html`

[12] Hunter, J. D., Matplotlib: A 2D graphics environment, Computing in Science & Engineering, 9:3 90–95, 2007, `https://matplotlib.org/`

[13] Matplotlib Navigation Toolbar, Matplotlib: A 2D graphics environment, `https://matplotlib.org/3.1.1/users/navigation_toolbar.html`

[14] M.A. Jaro. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, Volume 84, Issue 406, 1989.

[15] M.A. Jaro. Probabilistic linkage of large public health data files. *Statistics in Medicine*, Volume 14, Issue 5-7, 1995.

[16] W.E. Winkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Bureau of the Census*, 1990.

[17] V.I. Levenshtein. Binary Codes Capable Of Correcting Deletions, Insertions, and Reversals. *Cybernetics and Control Theory*, Vol 10, NO.8, 1966.

[18] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, Volume 48, Issue 3:443–453, 1970.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research , 12:2825–2830, 2011.

[20] Silhouette Score, Scikit-Learn, `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`

[21] Scikit-learn documentation for Violin Plots, Scikit-Learn, `https://scikit-learn.org/stable/modules/density.html`

[22] Textdistance. Textdistance. Python package, 2017, `https://pypi.org/project/textdistance/`.

[23] Bar Chart, Wikipedia, `https://en.wikipedia.org/wiki/Bar_chart`

[24] Cosine Similarity, Wikipedia, https://en.wikipedia.org/wiki/Cosine_similarity

[25] Damerau-Levenshtein Distance, Wikipedia, https://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance

[26] Directed Graph, Wikipedia, https://en.wikipedia.org/wiki/Directed_graph

[27] Exponential Distribution, Wikipedia, https://en.wikipedia.org/wiki/Exponential_distribution

[28] Floor and Ceiling Functions, Wikipedia, https://en.wikipedia.org/wiki/Floor_and_ceiling_functions

[29] Histogram, Wikipedia, https://en.wikipedia.org/wiki/Histogram

[30] Jaccard Index, Wikipedia, https://en.wikipedia.org/wiki/Jaccard_index

[31] Jaro Winkler Distance, Wikipedia, https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

[32] k-Medoids Clustering, Wikipedia, https://en.wikipedia.org/wiki/K-medoids

[33] Levenshtein Distance, Wikipedia, https://en.wikipedia.org/wiki/Levenshtein_distance

[34] Line Chart, Wikipedia, https://en.wikipedia.org/wiki/Line_chart

[35] Longest Common Subsequence, Wikipedia, https://en.wikipedia.org/wiki/Longest_common_subsequence_problem

[36] Needleman Wunsch algorithm, Wikipedia, https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

[37] Silhouette Score, Wikipedia, https://en.wikipedia.org/wiki/Silhouette_(clustering)

[38] Violin Plot, Wikipedia, https://en.wikipedia.org/wiki/Violin_plot